DOCUMENT RESUME

ED 442 854 TM 031 274

AUTHOR	Arenson, Ethan
TITLE	Estimating Error in State-to-NAEP Linkages, Part I: Mean Plausible Value Distributions from a Simple Model.
PUB DATE	2000-04-28
NOTE	38p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; *Equated Scores; Error of Measurement; *Estimation (Mathematics); National Competency Tests; Reading Tests; Regression (Statistics); State Programs; *Test Results; Testing Programs
IDENTIFIERS	Bootstrap Methods; Jackknifing Technique; *Linking Metrics; *National Assessment of Educational Progress

ABSTRACT

This paper is the first of a series that will compare estimates of error that arises when state assessments are linked to the National Assessment of Educational Progress (NAEP). Different forms of linkage are discussed. Comparisons are made between whole-sample regression, repeated half-sample replication, bootstrap, and jackknife estimates of the proportion of variance explained by the linkage function, as well as the standard error of linkage. Data from four states that participated in a study to link their state assessments to the 1998 (State) NAEP fourth and eighth grade reading assessments suggest that each of the methods produces comparable estimates of these error quantities when schools are treated as the primary sampling unit. (Contains 4 tables, 2 figures, and 13 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 442 854

Estimating Error in State-to-NAEP Linkages, Part I: Mean Plausible Value Distributions from a Simple Model

Ethan Arenson

School of Education
Stanford University
Stanford, CA 94305
arenson@stanford.edu

John C. Flanagan Research Center
American Institutes for Research
Palo Alto, CA 94304
earenson@ca.air.org

E. Arenson

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the American Educational Research Association, April 28, 2000, New Orleans, LA.

Last Revised: April 21, 2000

TM031274

BEST COPY AVAILABLE

Abstract

This is the first in a series of papers in which I attempt to compare estimates of the error that arise when one links state assessments to the National Assessment of Educational Progress (NAEP). A review of the different forms of linkages is discussed. Comparisons are made between whole-sample regression, repeated half-sample replication, bootstrap, and jackknife estimates of the proportion of variance explained by the linkage function, as well as the standard error of the linkage. Data from four states that participated in a study to link their state assessments to the 1998 (State) NAEP fourth and eighth grade reading assessments suggest that each of the methods produce comparable estimates of these error quantities, when schools are treated as the primary sampling unit.

Acknowledgements

In good conscience, I cannot make this presentation without recognizing the support I have received from my colleagues during the writing of this paper and of the other papers that have yet to be written. The fairest way I can express my gratitude is to do so alphabetically. I thank the following people for their encouraging words and constructive criticism: Larry Gallagher, Edward Haertel, Donald McLaughlin, Robert Mislevy, Ellen Reed, Nicole Schlesinger, Dale Whittington, and the four anonymous people who reviewed the proposal for this paper.

"Reporting individual student scores from the full array of state and commercial achievement tests on the NAEP scale and transforming individual scores on these various tests and assessments into the NAEP achievement levels are not feasible" (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; p. 91).

Introduction

"Linkage" refers to a process that translates scores from one test onto the scale of a second test. That is, linking facilitates a direct comparison of results from different tests (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). Linn (1993) mentioned that the need for linkages results from the demands that have been placed on educational testing:

"With increased demands for various types of assessments—from the classroom use of individual student results to international comparisons—has come an expanded desire to use assessments for multiple purposes by linking results from

distinct assessments. There is a desire to make comparisons on one assessment with those of another" (p. 83)[1].

One particular result from this increased demand has been a growing desire to link state assessments to the National Assessment of Educational Progress (NAEP). The desire for such linkages arises from both popular (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999) and political concerns (e.g., Goals 2000: Educate America Act, and current Title I legislation). In the case of the former, parents, schools, school districts, and states wish to compare their students against some common benchmark. The latter is merely legislation that requires states to compare their assessment results to a national benchmark. NAEP has assumed the role of this benchmark assessment.

If there were no financial, logistic, and legal constraints, there would be no need for linkages, as these agencies could administer NAEP (or some other benchmark assessment) to all students in their jurisdiction. Such a practice, however, is not feasible for several reasons, the strongest of which is that using NAEP for individual reporting purposes is illegal. Thus, some

[1] Until recently, the words *test* and *assessment* have been used interchangeably. The Joint Committee on Standards for Educational and Psychological Testing (1999) suggested the following distinction: *tests* measure "a sample of an examinee's behavior [e.g., academic performance] in a specified domain, whereas *assessments* integrate multiple sources of information, and may include results from multiple tests (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999).

form of linkage is needed that minimizes test burden (and, consequently, costs), yet also maximizes the amount of information available.

What is the National Assessment of Educational Progress?

Since the 1970s, NAEP has measured the educational achievement of young Americans in reading, writing, mathematics, science, U.S. history, and geography. Often referred to as "The Nation's Report Card," it accomplishes this task by collecting information on nationally representative samples of students who were 9, 13 and 17 years old (for the long-term trend sample); or in grades 4, 8 and 12 (for the main sample).

In each of the subject areas, NAEP estimates the achievement of a *group* of students as a composite score, based on a weighted average of subscales. Typical groups that NAEP uses for reporting include gender, ethnicity, and urbanicity (e.g., students from rural or urban areas). The distribution of performances are used to determine cut-points for achievement levels (basic, proficient and advanced). The National Assessment Governing Board (NAGB), the entity that oversees NAEP, emphasizes that achievement levels, not scores, are the primary way of

reporting NAEP results. Through the setting of achievement levels, NAGB can identify what students should know and should be able to do at various points on the NAEP scale.

The point that NAEP estimates achievement for groups of students cannot be overemphasized. NAEP employs a matrix-sampling technique, known as partially balanced incomplete-block (PBIB) spiraling. Through PBIB spiraling, each student receives a booklet that contains common blocks, consisting of background and motivational questions, in addition to other blocks that comprise the cognitive items. In the 1998 fourth- and eighth-grade reading assessments, each booklet contained two cognitive blocks, with the exception of one booklet in grade eight that consisted of one "extended" block. As a result of this design, no items in the assessment were common to all students. Arenson (1999) summarized the PBIB spiraling process for the 1998 NAEP reading assessment; full details can be found in Allen, Kline, and Zelenak (1986). Because each student answers a small portion of the items used in NAEP, it cannot accurately estimate individual proficiency.

Given that individual proficiencies are not known, NAEP estimates group proficiency by plausible values, scores randomly drawn from the student's "proficiency distribution." Plausible values incorporate significant variation to reflect the error due to sampling among students and with so few questions (Mislevy, 1991). For each student, NAEP generates five plausible values

per subscale. It is important to note that plausible values for a given student have no interpretation, other than serving as an intermediate step in determining the proficiency of subsamples of students.

NAEP calculates proficiency distributions by comparing, for each student in the subsample, the mean plausible value with the cutpoints for that grade level. The cutpoints for the basic, proficient and advanced proficiency levels on the 1998 reading assessment were the same as those used in the 1994 reading assessment: 208, 238 and 268 for grade four, and for grade eight 243, 281 and 323. For the subsample of interest, the proportions of students with mean plausible values falling in each of the four regions of the NAEP scale (below basic, basic, proficient, advanced) are estimated. It should be noted that NAEP uses sampling weights to determine the mean proficiencies and proficiency distributions for subgroups.

Methods for Linking Test Scores

There is common agreement in the literature that there are five methods of linking scores from distinct tests (Linn, 1993; Childs, 1996; Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). Listed in decreasing order of strictness, the methods are: equating, calibration, statistical

moderation, projection, and social moderation. In this section, I shall discuss, briefly, the differences among each of these methods. Detailed descriptions of these methods appear elsewhere (see, for example, Linn, 1993; Kolen & Brennan, 1995; Childs, 1996; Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). Table 1, taken from Linn (1993), summarizes the differences discussed below.

INSERT TABLE 1 ABOUT HERE

Equating

Equating, the most rigorous of linkage methods, is typically reserved for the linking of scores from tests that one has intentionally designed to be parallel. That is, tests to be equated must measure the same construct and must be equally reliable. This requirement is implicit in Lord's (1980) equity requirement:

> "If an equating of tests x and y is to be equitable to each applicant, it must be a *matter of indifference* to applicants at every given ability level θ whether they are to take test x or test y" (p. 195, italics added).

The equating function adjusts scores from one test (e.g., test x) so that one may compare them with scores from a second test (test y). A strict interpretation of Lord's equity requirement leads to the conclusion that equating tests is a nonsensical exercise, for if two tests satisfy the equity requirement, then by definition they are interchangeable and their scores do not need to be equated. In common practice, however, this requirement is relaxed.

Another requirement for the equating of scores from distinct tests is that the equating function be population invariant (P. W. Holland, personal communication, 19 January 2000). That is, there should be a "matter of indifference" regarding the choice of sub-population used to compute the equating function. Whereas Lord's equity requirement demonstrates indifference among tests, population invariance demonstrates indifference among subsets of the population. There are additional conditions that must be satisfied before one can justify the linking of test scores by equating; these conditions are presented in greater detail elsewhere (e.g., Kolen & Brennan, 1995; Holland & Rubin, 1982).

It must be stated at the outset that state tests and NAEP cannot be equated. One important reason establishing the inappropriateness of equating is that the two tests cannot measure the same construct. The NAEP framework is typically broader than the framework for any particular state. As such, a considerable degree of mismatch between the two tests is inevitable. One can ultimately determine the degree of mismatch by analyzing the content of both tests. While state tests may be available for content analyses, NAEP tests are not. NAEP maintains tight security over its booklets, in order to recycle items for subsequent administrations.

Calibration

Calibration is the second-most stringent form of linkage. In essence, one calibrates two tests when the desire is to compare scores (typically from a short form) of one test to scores from another, generally longer, test. Linn (1993) illustrated calibration with the following example:

> "A state uses a version of a test that is shorter than a national test but designed to measure the same skills. The state version is less reliable than the national test due to its reduced length. Estimates of the percentage of students in the state who score above selected points on the national test are desired" (p. 86).

Calibration does not need to be population invariant. That is, calibrations may differ for different sub-populations. However, the tests do need to measure the same construct. For this reason, a calibration of state tests with NAEP, for reasons previously mentioned, is not appropriate.

Statistical Moderation

In terms of stringency, the next method of linkage is statistical moderation. Though statistical moderation is common in countries outside the United States, the American equivalent involves the use of an anchor test (Linn, 1993; Childs, 1996). In either case, the objective is to adjust scores from a "locally-administered" test in a way that sets the mean and variance of the scores on the local test equal to the within-school mean and variance of scores on an external test (i.e., an anchor test).

Statistical moderation is not appropriate for state-to-NAEP linkages, because statistical moderation assumes a different relationship between the two tests. As hypothetical scenario for which statistical moderation is appropriate, consider a state that wishes to use school-level NAEP scores (the external test) to adjust individual scores on state tests (the local test). Putting scaling

considerations aside, such a situation is not realistic, since only a few schools in a given state participate in NAEP.

Social Moderation

Social moderation is a method of linkage that is consensus-based, rather than statistical in nature. That is, test scores are "moderated" by judges trained to make scores from one test comparable to those on another test. Typically, this method is used to compare ratings of student products, such as performance tasks. Similar to the way in which trained readers would grade essays, this method requires an agreed-upon set of standards and exemplars of student work along the continuum of student performance. Linn (1993) gave the following example:

> "States or groups of states develop their own sets of performance-based assessments in reference to a common content framework. Scoring of performances depends heavily on professional judgments of teachers and a system of spot checks and verification. Nonetheless, it is expected that performance of individual students, schools, districts, and states will be compared to a single set of national standards" (p. 87).

By design NAEP and practically all state tests do not lend themselves to linking by social moderation.

Projection

The least statistically rigorous form of linkage is projection. Also referred to as prediction, this method uses information from one or more tests—such as scores and, on occasion, demographic data—to predict performance on another test. Projection does not require that the tests measure the same construct, nor does it require that the tests be equally reliable. In fact, McLaughlin (1998) commented that one could create a statistically sound projection to predict I.Q. from hair length. The statistical soundness of a projection does not guarantee that it will be conceptually helpful.

Projection is appropriate for linking state tests to NAEP, and is the method used for this study. This method is discussed in greater detail in the Methodology section.

Regardless of the method one uses, linkage studies between state tests and NAEP typically address two questions. *First*, how well do state tests predict mean NAEP plausible

values for students in certain subpopulations? *Second*, how well do state tests predict the proficiency distribution (i.e., percent of students performing at the basic, proficient, and advanced levels) for specific subpopulations?

This paper is the first of three papers in which I attempt to quantify the amount of error that arises when state test scores are linked to the NAEP scale. In particular, I compare the estimates of the proportion of variance (i.e., the adjusted squared correlation) and of the linkage error based on simple projection models (explained in more detail in the Methodology section) used to predict mean NAEP plausible values from state test scores and demographic data. The second paper will compare methods of estimating errors based on the misclassifications of proficiencies for groups of students. The final paper will revisit these two explorations under a generalized linkage model.

Before discussing the methodology of this study, I must clarify an important distinction. This paper is the product of a current linkage study in which state test scores and demographic variables are used to predict 1998 mean NAEP plausible values and proficiency distributions for fourth and eighth grade students who both took the state test(s) and participated in NAEP (i.e., those who have NAEP plausible values). The purpose of this paper is not to report the results of the linkage study; these results will appear in a future report. The sole function of these

preliminary models is to permit comparisons of methods that estimate the adjusted squared correlations and the root-mean-squared errors for each state and grade level combination.

Methodology

Let ρ^2 denote the true proportion of variance in NAEP plausible values that is explained by the linkage model (i.e., the true adjusted squared correlation). Similarly, let σ denote the true root-mean-squared error for this model. These parameters will be estimated by four methods: whole-sample regression, half-sample replication, bootstrap and the jackknife. The respective statistics that estimate these quantities shall be denoted by r^2_{whole}, $RMSE_{\text{whole}}$, r^2_{half}, $RMSE_{\text{half}}$, r^2_{boot}, $RMSE_{\text{boot}}$, r^2_{jack}, and $RMSE_{\text{jack}}$.

The Sample

Data come from four states that participated in a study to determine the feasibility of linking 1998 state reading test scores to the 1998 NAEP reading scale.[2] The linkage study sample consists of students who took their state's test(s) and who participated in NAEP (i.e., students

who have NAEP plausible values). Three states provided data for both fourth- and eighth-grade students; one state provided data for only fourth grade students. Table 2 shows the numbers of students and of schools at each grade for each state that participated in the linkage study.

INSERT TABLE 2 ABOUT HERE

The sizes of the linkage samples among the state-grade combinations ranged between nearly 1,650 and 2,400 students from roughly between 50 and 110 schools.

Whole-Sample Regression

For each state-grade combination, four multiple linear regressions were computed and compared: whole sample (reverse stepwise), half-sample, bootstrap, and jackknife. Reverse stepwise multiple linear regression was used to determine which variables would determine the

[2] Confidentiality agreements require anonymity among the participating states.

benchmark model against which the error estimates from other models would be compared. The full array of variables used in the first step of the regression consisted of standardized test scores[5], squared values of these standardized scores, ethnicity (measured as white or minority), gender, school-level mean test scores, and school percentages of minority and female students. For some state-grade combinations a ceiling effect on NAEP was present. It is to compensate for this ceiling effect that quadratic terms were included in the initial model. The full model can be written as

$$\mathbf{y} = \mathbf{s}\hat{\boldsymbol{\alpha}} + \mathbf{q}\hat{\boldsymbol{\beta}} + \mathbf{d}\hat{\boldsymbol{\gamma}} + \mathbf{g}\hat{\boldsymbol{\delta}} + \boldsymbol{\varepsilon}, \qquad (1)$$

where, for a state with n_{stu} students and k tests in the model, $\mathbf{y}$ denotes an $n_{stu} \times 1$ matrix of NAEP plausible values, $\mathbf{s}$ denotes an $n_{stu} \times k$ matrix of k test scores for each student, $\mathbf{q}$ denotes an $n_{stu} \times k$ matrix of quadratic terms for each of the k tests, $\mathbf{d}$ denotes an $n_{stu} \times 2$ matrix of demographic information (ethnicity and gender), $\mathbf{g}$ denotes an $n_{stu} \times (k+2)$ matrix containing school-level means for each test and school-level percentages of the demographic variables, and $\boldsymbol{\varepsilon}$ denotes an $n_{stu} \times 1$ matrix of residuals. $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$, and $\hat{\boldsymbol{\delta}}$ denote vectors of the estimated regression coefficients for each of the respective matrices, $\mathbf{s}$, $\mathbf{q}$, $\mathbf{d}$, and $\mathbf{g}$. As a matter of

[5] Standardized test scores were used to preserve anonymity.

convenience, let $\mathbf{X}$ denote the $n_{stu} \times (3k + 5)$ matrix that contains all the elements of matrices $\mathbf{s}$, $\mathbf{q}$, $\bar{\mathbf{d}}$, and $\mathbf{g}$, and let $\mathbf{b}$ denote the vector of corresponding estimated coefficients, including the intercept. Then, (1) can be written more succinctly as

$$\mathbf{y} = \mathbf{X}\hat{\mathbf{b}} + \boldsymbol{\varepsilon}. \tag{2}$$

where $\mathbf{X}$ is of dimensionality $n_{stu} \times k'$, with k' denoting the $3k + 5$ coefficients. Similarly, $\hat{\mathbf{b}}$ is a k'-dimensional vector. For simplicity, it is assumed that both measurement error and error from estimating the regression coefficients are subsumed in the error vector $\boldsymbol{\varepsilon}$; the third paper in this series will address the complications that arise when one explicitly includes measurement error and error from estimating regression coefficients.

Computing r^2_{whole} and $RMSE_{whole}$ require calculating the linkage, error, and total sums of squares. If $\mathbf{J}$ denotes an $n_{stu} \times n_{stu}$ matrix with the value "1" in every element, then

$$SS_{link} = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \left(\frac{\mathbf{1}}{n_{stu}}\right)\mathbf{y}'\mathbf{J}\mathbf{y} \tag{3}$$

$$SS_{error} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \tag{4}$$

$$SS_{\text{total}} = \mathbf{y}'\mathbf{y} - \left(\frac{\mathbf{1}}{n_{\text{stu}}}\right)\mathbf{y}'\mathbf{J}\mathbf{y}. \tag{5}$$

Given the sums of squares above, one can compute r^2_{whole} by Equation (6) below:

$$r^2_{\text{whole}} = \left(\frac{n_{\text{stu}} - 1}{k' - 1}\right)\frac{SS_{\text{link}}}{SS_{\text{total}}} \tag{6}$$

Note that the mean-squared error (MSE_{whole}) is merely $\frac{SS_{\text{error}}}{n_{\text{stu}} - k}$, and that $RMSE_{\text{whole}}$ is the square root of this quotient.

In determining the benchmark model, the least statistically significant variable was removed from the model, and a new regression model was determined. This process of removing variables continued iteratively until all variables were statistically significant at the .05 level. From the final regression model, point estimates for r^2_{whole} and $RMSE_{\text{whole}}$ were obtained. These point estimates would be compared to the confidence intervals for these statistics, as estimated from the other methods.

Half-Sample Replications

In addition to obtaining point estimates from whole-sample regression models, error estimates were obtained by averaging the estimates from 100 half-sample replications. For each iteration, each school in the linkage sample was randomly assigned to either a training sample or a validation sample. That is, the training and validation samples were selected without replacement from the linkage sample. For each replicate, test scores and demographic indicators from the training sample were used to predict the sum of the mean NAEP plausible value and a random error component. The addition of random follows the methodology that McLaughlin (1998) has used previously for predicting group-level NAEP proficiency distributions. Let $r^2_{i_{\text{half}}}$ and $RMSE_{i_{\text{half}}}$ denote the adjusted r-squared and the root-mean-squared error for the training sample of the ith half-sample replicate. Then, if $E(\cdot)$ denotes the expected value; and—for the ith replicate—$SS_{i_{\text{link}}}$ denotes the sum of the squared deviations for the linkage function, and $SS_{i_{\text{total}}}$ denotes the sum of the squared deviations among the NAEP plausible values,

$$r^2_{\text{half}} = E\left(r^2_{i_{\text{half}}}\right) = E\left(\frac{SS_{i_{\text{link}}}}{SS_{i_{\text{total}}}}\right), \text{ and} \tag{7}$$

$$RMSE_{\text{half}} = E\left(RMSE_{i_{\text{half}}}\right) = E\sqrt{MSE_{i_{\text{half}}}}\,. \tag{8}$$

95-percent confidence intervals were constructed for each point estimate. For simplicity, let t denote the statistic of interest (r^2_{bolt} or $RMSE_{\text{bolt}}$). The 95-percent confidence interval is determined by

$$t_{\text{bolt}} \pm 1.96\sqrt{Var(t_{\text{bolt}})}. \tag{9}$$

where $Var(\cdot)$ denotes the variance.

Bootstrap

The bootstrap (Efron & Tibshirani, 1993) was also used to estimate the error parameters of interest. Let n_{sch} denote the number of schools in a state's linkage sample. 1,000 random samples with replacement of n_{sch} schools each were selected as pseudo-samples. The benchmark linkage model (based on the whole-sample regression mentioned previously) was applied to each pseudo-sample, and the estimates $r^2_{\text{bootstrap}}$ and $RMSE_{\text{bootstrap}}$ were determined, using formulae similar to Equations (7) and (8):

$$r^2_{boot} = E(r^2_{boot}) = E\left(\frac{SS_{boot}}{SS_{boot}}\right) \quad (10)$$

$$RMSE_{boot} = E(RMSE_{boot}) = E\sqrt{MSE_{boot}} \quad (11)$$

Similar to Equation (9), 95-percent confidence intervals were constructed for the point estimates r^2_{boot} and $RMSE_{boot}$ were determined by

$$t_{boot} \pm 1.96\sqrt{Var(t_{boot})}. \quad (12)$$

The Jackknife

The jackknife (Efron & Tibshirani, 1993; Mosteller & Tukey, 1977) was the final method for estimating error parameters in the state-to-NAEP linkages. In each state, n_{sch} pseudo-values were calculated by deleting observations from school i in the ith pseudo-sample $(i = 1,2,\ldots,n_{sch})$. Specifically, let $r^2_{(i)}$ and $MSE_{(i)}$ denote the proportion of variance explained by the linkage and the mean-squared error with the ith school deleted. Similarly, let r^2_{all}, and MSE_{all} denote the corresponding statistics for the linkage sample. Then,

$$MSE_{(j)} = n_{sch} MSE_{all} - (n_{sch} - 1) MSE_{(j)}, \text{ and} \tag{13}$$

$$\left(r_{(j)}^{2}\right)^{2} = n_{sch} r_{all}^{2} - (n_{sch} - 1) r_{(j)}^{2}. \tag{14}$$

Similar to Equations (7) and (8), point estimates from the jackknife are denoted by

$$r_{jack} = E\left[\left(r_{(j)}^{2}\right)^{2}\right] = E\left(\frac{SS_{t_{jack}}}{SS_{t_{jack}}}\right) \tag{15}$$

$$RMSE_{jack} = E\left(RMSE_{t_{jack}}^{2}\right) = E\sqrt{MSE_{t_{jack}}^{2}}. \tag{16}$$

95-percent confidence intervals were determined by

$$t_{jack} \pm 1.96\sqrt{\frac{t_{jack}}{n_{sch}-1}}. \tag{17}$$

Results

As shown in Tables 3 and 4, each of the error estimation methods—whole-sample regression, half-sample, bootstrap, and jackknife—produced comparable point-estimates for the parameters ρ^2 and σ. Furthermore, the half-sample, bootstrap, and jackknife methods produced similar degrees of variability in these estimates. That is, the standard errors were similar.

INSERT TABLES 3 AND 4 ABOUT HERE

Table 3 also shows between-state variation in the point-estimates. This variation is reflected in the six panels of Figure 1 (estimates of ρ^2) and Figure 2 (estimates of σ).

INSERT FIGURES 1 & 2 ABOUT HERE

Discussion

Tables 3 and 4, as well as Figures 1 and 2, show considerable variation in estimates of error between states and grades. This variation is to be expected, given the specific tests that each state chose to administer. That is, different tests will have varying degrees of content overlap with NAEP. A test that correlated perfectly with NAEP, assuming that such a test existed, would have the least amount of error. Conversely, a test that had no content overlap with NAEP (i.e., the two tests were uncorrelated) would have the largest amount of error. It is worth noting that in the case of a test that correlated perfectly with NAEP, there would still be some linkage error. The linkage error, in this case, reflects the inherent error due to NAEP sampling and to plausible value estimation (McLaughlin, 1998).

Within each state-by-grade combination, the estimates of error from each of the different estimation methods produce similar results. The 95-percent confidence intervals for the half-sample, bootstrap and jackknife methods all capture the estimates of ρ^2 and σ that were obtained by whole-sample regression.

The results from this study constitute a first-step towards developing a better understanding of the error that arises when state tests are linked to NAEP. Other studies are

clearly needed to answer unanswered questions. *First*, how similar are the four error estimation methods presented here in estimating proficiency distributions? This question is particularly important, given the emphasis on reporting NAEP results in terms of proficiency distributions (Pellegrino, Jones, & Mitchell, 1999).

Second, would a more complex linkage model that incorporates test reliabilities produce more accurate estimates of error? The models in this study were based on the premise that errors of measurement were subsumed under one (general) error term. A related question is the following: will the uses of more reliable tests in a linkage improve error estimates, or is linkage error solely an artifact of the correlation between the tests linked to NAEP?

Third, do the linkage models overspecify the relationship between the predictor test scores and the mean NAEP plausible values? Linkage models that include minority or gender indicator variables are, in a sense, redundant, because NAEP plausible values are derived by conditioning on the first 200 principal components of a variety of demographic and background, including ethnicity and gender. The issue of whether a model that links state test scores and demographic variables to mean NAEP plausible values that are *not* conditioned on demographic variables merits further exploration.

Conclusion

This study compared whole-sample, half-sample, bootstrap, and jackknife methods of estimating the linkage error and the proportion of variance explained by the linkage model for fourth and eighth grade students from selected states that participated in the NAEP 1998 reading assessment. Treating schools as primary sampling units, the point-estimates of error and their standard errors obtained by whole-sample regression were comparable to the estimates obtained by half-sample, bootstrap and jackknife methods. A comparison of methods of estimating error in predicted proficiency distributions and a study of whether knowledge of test reliabilities can improve these error estimates will be topics for subsequent papers.

References

Allen, N. L., Johnson, E. G., Mislevy, R. J., & Thomas, N. (1996). Scaling procedures. In N. L. Allen, D. L. Kline, & C. A. Zelenak (Eds.), *The NAEP 1994 technical report*. Washington, D. C.: National Center for Educational Statistics (Publication No. 97-897).

Allen, N. L., Kline, D. L., & Zelenak, C. A. (1996). *The NAEP 1996 technical report*. Washington, D. C.: National Center for Educational Statistics (Publication No. 97-897).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards on educational and psychological testing*. Washington, D. C.: American Psychological Association.

Arenson, E. A. (March, 1999). Statistical linkages between state education assessments and the National Assessment of Educational Progress. Paper presented at the annual meeting of the Sacramento Statistical Association, Sacramento, CA.

Childs, R. (1996). Review of the literature on linking state and national assessments. Unpublished manuscript. Washington, D. C.: American Institutes for Research.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York: Chapman & Hall.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, D. C.: National Academy Press.

Holland, P. W., & Rubin, D. B. (1982). *Test equating*. New York: Academic Press.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating*. New York: Springer-Verlag.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.

McLaughlin, D. (1998). Study of the linkages of 1996 NAEP and state mathematics assessments in four states. Palo Alto, CA: American Institutes for Research.

Mosteller, F. & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Table 1. Requirements of different linkage methods.

	Method of Linkage				
Requirement	Equating	Calib.	Stat. Mod.	Proj.	Soc. Mod.
Measure the same construct	Yes	Yes	No	No	No
Be equally reliable	Yes	No	No	No	No
Use an external anchor	No	No	Yes	No	No
Be population invariant	Yes	No	Yes	No	Yes
Use exemplars of performance	No	No	No	No	Yes
Use trained judges	No	No	No	No	Yes

From "Linking Results from Distinct Assessments" by R. L. Linn, 1993, *Applied Measurement in Education, 6*, 83-102.

Table 2. Descriptive information for states participating in the linkage sample.

	Grade 4		Grade 8	
State	$N_{students}$	$N_{schools}$	$N_{students}$	$N_{schools}$
A	2,196	106	1,889	49
B	2,384	107	2,380	104
C	1,647	90	1,696	86
D	2,250	95		

Note: Confidentiality agreements require anonymity among the participating states.

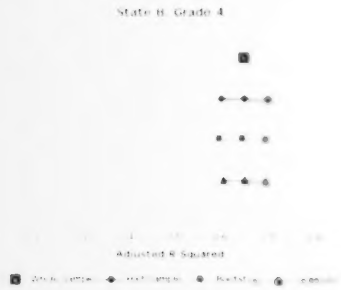
Table 3. Point estimates and standard errors of adjusted *r*-squared values by various error estimation methods.

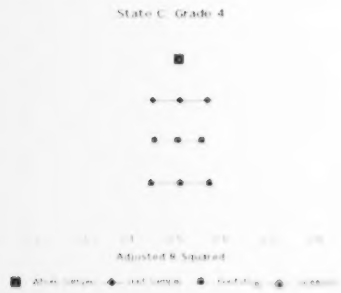
Grade 4			Grade 8		
State A	Est.	SE	State A	Est.	SE
Whole-Sample	0.549		Whole-Sample	0.442	
Half-Sample	0.542	0.022	Half-Sample	0.438	0.027
Bootstrap	0.552	0.023	Bootstrap	0.418	0.020
Jackknife	0.549	0.025	Jackknife	0.442	0.031
State B	Est.	SE	State B	Est.	SE
Whole-Sample	0.652		Whole-Sample	0.548	
Half-Sample	0.654	0.025	Half-Sample	0.547	0.030
Bootstrap	0.649	0.025	Bootstrap	0.546	0.026
Jackknife	0.653	0.023	Jackknife	0.548	0.032
State C	Est.	SE	State C	Est.	SE
Whole-Sample	0.513		Whole-Sample	0.429	
Half-Sample	0.514	0.030	Half-Sample	0.436	0.016
Bootstrap	0.511	0.026	Bootstrap	0.430	0.034
Jackknife	0.514	0.032	Jackknife	0.428	0.024
State D	Est.	SE			
Whole-Sample	0.520				
Half-Sample	0.479	0.029			
Bootstrap	0.523	0.021			
Jackknife	0.518	0.026			

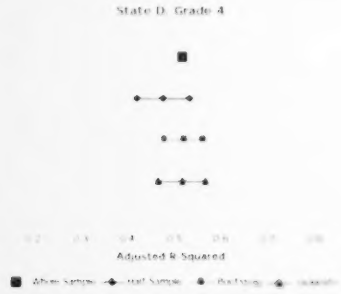
Table 4. Point estimates and standard errors of linkage errors by various error estimation methods.

Grade 4			Grade 8		
State A	Est.	SE	State A	Est.	SE
Whole-Sample	21.00		Whole-Sample	20.68	
Half-Sample	21.11	0.50	Half-Sample	20.82	0.45
Bootstrap	20.76	0.43	Bootstrap	20.72	0.45
Jackknife	21.05	0.46	Jackknife	20.75	0.51
State B	Est.	SE	State B	Est.	SE
Whole-Sample	17.47		Whole-Sample	19.55	
Half-Sample	17.40	0.47	Half-Sample	19.54	0.78
Bootstrap	17.36	0.48	Bootstrap	19.46	0.49
Jackknife	17.51	0.45	Jackknife	19.59	0.80
State C	Est.	SE	State C	Est.	SE
Whole-Sample	25.01		Whole-Sample	24.18	
Half-Sample	24.68	0.54	Half-Sample	24.20	0.09
Bootstrap	24.79	0.66	Bootstrap	24.00	0.50
Jackknife	25.06	0.66	Jackknife	24.24	0.46
State D	Est.	SE			
Whole-Sample	21.60				
Half-Sample	22.19	0.81			
Bootstrap	21.34	0.44			
Jackknife	21.73	0.59			

Figure 1. Estimates of adjusted *r*-squared values by various methods.

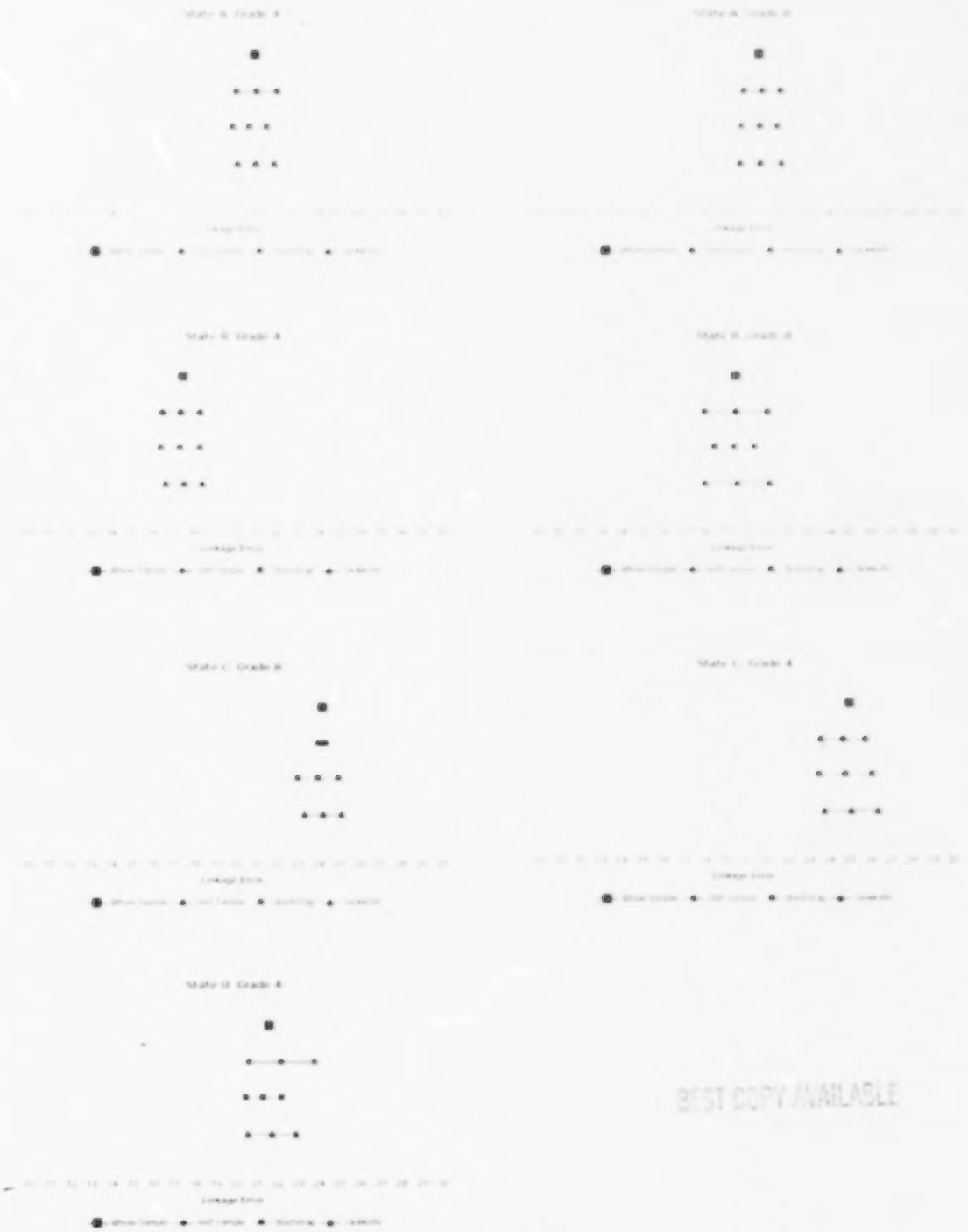






BEST COPY AVAILABLE

Figure 2. Estimates of linkage errors by various methods.



BEST COPY AVAILABLE